微博舆情传播周期中不同传播者的主题挖掘与观点识别*

廖海涵[1] 王曰芬[1,2] 关鹏[1]
[1] 南京理工大学经济管理学院 南京 210094 [2] 江苏省社会公共安全科技协同创新中心 南京 210094

摘要：[目的/意义]探索微博舆情传播周期中不同传播者关注的舆情热点和传播内容的主要观点，进而发现舆情传播的特点和规律，为舆情分析与决策提供依据。[方法/过程]以特定舆情事件的事实文本数据为来源，以生命周期理论和 LDA 方法为指导，设计研究流程与构建研究模型，对微博舆情事件中不同传播者的话题进行主题研究，其中包括主题抽取和结果语义标注、各阶段的不同传播者主题的语义分析、基于时间维度的舆情主题观点识别与刻画。[结果/结论]研究发现，论文所提出的研究模型能够挖掘出舆情传播周期中不同传播者的主题结构、观点脉络以及特征，研判出分布在文字当中有关联性的、代表性的、重要的词语。同时，结论中还发现微博中的官媒、大众媒体发布信息中的话题和用户谈论的热点话题具有明显的差异性。

关键词：微博舆情 不同传播者 主题挖掘 观点识别 生命周期理论 LDA 主题模型
分类号：G250
DOI：10.13266/j.issn.0252-3116.2018.19.010

1 引言

新媒体催生出社会舆情传播形式的多元化、传播速度的瞬时化、传播数量的海量化、传播内容的高度分散化与各种会话的碎片化，常常使得社会事件的新闻报道在发布与评论过程中不断地被放大或者被扭曲，往往造成社会舆情事件的频繁爆发。其中，以微博为代表的新媒体，不仅因其具有的匿名性、自由性等特点使网民任性表达的意愿得到激发，而且由于交流功能的便捷性与社交性使得传播元素之间的相互作用得到彰显。经由微博发布的舆情事件，传播后产生的话题内容褒贬不一、传播者交互的情感和态度多元复杂与影响多样，强化了舆情传播的突发性与效应性，增加了社会不稳定因素与复杂程度，加大了公共治理的难度。那么，在社会舆情传播的不同阶段，信息发布后，经过微博传播产生的信息评论，在主题内容上是如何发展并形成哪些表达观点；同时，与信息发布主题相比，舆情评论主题又有何异同。由此，本文从舆情监测实时性的视角出发，结合生命周期理论、LDA 模型，构建一种有效地动态挖掘舆情事件热点主题的分析模型，并依据研究模型追踪舆情传播内容，挖掘出隐藏在舆情传播阶段万千话语中的热点主题，以探索时间维度下信息发布者与信息接收者的两类不同传播者的语言特征和表达观点。

2 相关工作

根据研究的需要，本文首先对舆情主题研究现状的相关文献加以综述；然后，再对研究使用方法和理论加以梳理。

2.1 舆情主题研究现状

通过文献调研，国内外学者对于舆情主题的研究主要从舆情主题挖掘和主题监测的视角进行展开。在主题挖掘研究中常应用自然语言处理、文本聚类、共词分析、主题建模、算法改进等技术方法。在主题监测层面中，主要从主题监测追踪、预警等研究角度进行研究。

在主题挖掘研究方面，学者陈晓美等基于 LDA 主题模型观点提取方法，通过对比分析观点提取方法间的差异，从认知上阐释网络舆论平台的群体智慧和受众个体的认知过程，最后发现 LDA 主题模型提取舆情观点的优势及新路径[1]。张寿华等在舆情热点主题研究中采用了 TFIDF 法、话题聚类算法等方法，研究还设

* 本文系国家社会科学基金重点项目"大数据环境下社会舆情与决策支持方法体系研究"（项目编号：14AZD084）研究成果之一。
作者简介：廖海涵（ORCID：0000-0003-4953-1075），博士研究生；王曰芬（ORCID：0000-0002-7143-7766），教授，博士生导师，通讯作者，E-mail：yuefen163@163.com；关鹏（ORCID：0000-0002-2308-3019），副教授，博士。
收稿日期：2018-03-12 修回日期：2018-06-11 本文起止页码：77-85 本文责任编辑：杜杏叶

计了一个主题挖掘系统,并通过舆情预处理、关键词提取、话题聚类、热点话题分析等关键环节来实现,研究发现所设计的系统对网络舆情热点有较高的识别准确率[2]。李磊等应用共词分析方法对舆情主题进行研究,并在研究中通过构建共词矩阵、关键词共现找出热点话题,同时研究结果表明提出的研究方法不仅有应用价值,且可提高对网络舆情信息精炼和概括的效率[3]。钱爱兵应用如主题关注、热点、焦点等舆情信息的计量方法,通过构建基于主题的网络舆情分析模型,包括:舆情主题规划、舆情信息采集和分析、舆情预警等,得到舆情主题关注、热点、焦点、重点的研究结论[4]。梁晓贺等通过网络计量方法,在研究中构建用户、观点、情感、时序阶段4层子网的超网络模型,并将该模型结合具体案例事件进行分析,其研究结论表明,基于舆情主题发现的超网络模型的子网分析可揭示每层子网的特征信息,超边分析可用于舆情预警分析、舆情主题挖掘及舆情主题演化分析[5]。N. Li等通过K-means聚类和SVM算法对文本进行挖掘研究,将新浪微博的体育论坛进行归类,并进行文本挖掘,研究结论挖掘出基于文本数据的热门话题,和发现两种方法的应用都得到了同样的结果[6]。L. Y. F. Su等基于HK算法研究舆情主题,通过智能算法改进、内容分析,对传播主题情感进行挖掘,结果发现该研究方法对社交媒体舆情主题挖掘的研究具有可靠性和有效性[7]。

在舆情主题监测研究方面,丁晟春等采用了网络爬虫、网页预处理、正文提取等技术方法,通过构建针对南海问题的多语种舆情监测体系,并实行时间序列下的主题追踪,研究结论表明构建的舆情监测系统能够实现舆情信息采集、处理和分析等[8]。张瑜等基于巴斯模型,通过微博文本分词、阈值确定、微博文本特征提取、议题词典构建、微博文本议题划分等文本处理步骤,实现对微博热点事件内部不同主题的情感随时间分布下变化与发展趋势的研究,结果发现了不同话题的状态、群体情感对话题的影响等结果[9]。安璐等采用了生命周期理论和word2vec技术,通过对主题评论情感做细粒度划分,计算情感强度,最终实现微博主题与情感的协同分析,结果发现研究所提出的分析方法能够揭示面向特定事件的微博网络舆情主题与情感特征的协同演化规律[10]。J. Zhao等应用NB算法,并通过改进算法、定义参数,对舆情非常态事件进行情感计量,研究达到了时间序列下舆情监测的意义[11]。Q. Mei等应用TSM模型对微博主题进行监测,通过模型中的参数定义建立研究模型,再应用生命周期划分对研究进行实证,研究发现建立的模型有很好的挖掘和监测效果以及潜在的应用价值[12]。

综上所述,舆情主题研究是许多研究者关注的内容,研究数量不断增加,并形成许多有价值的研究成果和观点。根据上述研究还可知,关于网络舆情的主题挖掘研究多基于文本挖掘技术和智能算法等方法来实现;在主题挖掘中,自然语言处理的LDA模型、特征词抽取等方法能够较精准的揭示语料库的词语特征、适合大规模数据集的挖掘研究;而共词分析法,存在词频阈值的不确定性;和主题词的社会网络分析法,注重词语间的关联性,适合应用于小数据集的分析[13];文本聚类方法中的标注类别具有较强的主观性,也比较适合少量的数据研究。主题监测方面大多数研究都会基于时间动态视角进行追踪研究,往往涉及到生命周期理论、时间序列等理论与方法;在技术方法上涉及到word2vec技术,NB算法、TSM模型、统计分析等方法;通常生命周期等理论通过与技术方法的结合应用来展现监测主题的变化、变异等形态和趋势。但是现今舆情主题监测研究中,大多数缺乏对不同性质主题的辨别,使得监测主题的内容与观点散乱、无着力点。由于本文的研究即将进行较大规模数据处理,且力求发现时间粒下的不同性质传播者舆情主题的内容和观点。所以,本文采用LDA模型方法和生命周期理论,以两者的结合应用为指导,提出研究流程和研究模型,以揭示网络舆情发展的规律。

2.2 研究方法

2.2.1 LDA主题模型　LDA主题词挖掘是自然语言处理中的重要挖掘方法,也是一种完全生产式的模型。LDA主题模型可以展现出单个主题下相关词项的集合及概率,能够排除主观因素对于科学研究的影响,还弥补传统研究不能有效深入挖掘大批量文本的局限。由于,LDA模型在国内外研究中的应用已相当广泛和成熟,研究不再赘述。

在舆情主题研究中,唐晓波等基于LDA模型对微博热点进行挖掘,该研究构建了关于微博热度概念的LDA模型,然后通过采集的微博数据进行实验,研究结果发现改进的LDA模型能得到更直观的微博热度表和更具有说服力的挖掘结论[14];林萍等基于LDA模型抽取话题,通过后离散时间型话题模型思路分析话题热度变化,和先离散时间型话题模型思路分析话题内容迁移,研究不仅发现舆情事件话题内容,还发现最佳话题数量与文本内容焦点集中度密切相关等结论[15];W. X. Zhao等采用LDA文本挖掘技术,对Twitter的内

容和传统媒体纽约时报进行同类主题的挖掘对比，深入探究了发布与回复博文的主题和类别之间的关系，研究还发现线下和线上的异同情况[16]；M. Pennacchiotti 等采用 LDA 主题模型，发现用户的兴趣，最终研究得出了一个向用户推荐相似兴趣朋友的系统[17]。综上所述，LDA 主题模型不仅是舆情主题分析的一种有效方法，也是学者们倾向采用的热点技术。

2.2.2 生命周期理论 生命周期理论能够很好的揭示事物从诞生、生长、成熟、衰退到消亡的过程，在各学科领域都得到了广泛的应用。在网络舆情的研究中，一般将信息从产生到失效的整个经历过程定义为传播的生命周期。生命周期理论作为一种概念理论，需要具体的环节进行支撑，其应用相对于模型研究、仿真模拟等方法有更进一步的现实意义。

关于生命周期理论在舆情主题中的研究有：安璐等利用 SOM 自组织映射、生命周期理论等方法，通过文本预处理、主题分类等步骤，对 Twitter 与微博平台上关于西非埃博拉病毒爆发的热点主题进行对比分析，研究结论发现了主题演化模式和时序趋势的异同点[18]。陈福集等采用了话题传播演化博弈模型，通过生命周期的情景预测与模型拟合，对话题传播做出相应的研究，并且研究还总结了基于演化博弈的网络舆情热点话题传播模型的对策[19]。张思龙在其研究中借鉴“微博生命周期”理论，设计了基于“微博关注度”的话题多元信息动态更新机制[20]。Q. Mei 等采用了一种新的概率方法来构建舆情主题研究模型，并且结合生命周期理论，通过主题生命周期的划分，对每一个给定的时间段主题生成快照，研究结果表明所构建的研究模型能够适用于普遍的时间和空间信息的分析[21]。那么，由上述研究的特点可发现，以生命周期传播规律理念为指导的研究可更深入、细致地挖掘出舆情主题传播规律中的有效信息和结论。

3 研究设计

LDA 模型能够准确、清晰地表达主题中隐藏的信息，而生命周期理论则能够从微观上展现微博传播主题在时间粒下的变化细节、特征、特点等。由此，本文基于 LDA 与生命周期理论提出本文的研究设计，设计中包括：研究思路与流程，生命周期理论的划分规则、不同传播者的热点主题模型构建 3 个环节。

3.1 研究思路与流程

为实现舆情传播周期阶段中不同传播者的热点主题挖掘与观点识别，研究的思路设计为：首先，对微博舆情事件数据进行归一化处理；其次，以生命周期理论为依据划分舆情传播周期的不同阶段；再次，基于 LDA 模型进行主题挖掘与观点识别；最后，进行词频统计，并与前述的主题挖掘结果对比，以验证本研究思路与结论。具体的研究流程及设想描述如下：

3.1.1 微博文本数据归一化 微博中的异构数据会直接影响到主题抽取的结果，那么，本研究先要解决的就是微博异构数据归一化，即将所有异构类型的数据清理，统一转换成规范的数据格式保存，以为主题抽取和语义挖掘做基础铺垫。归一化的方法涉及切词、分词、停用词过滤等自然语言处理过程。

3.1.2 基于生命周期理论的舆情事件传播周期阶段的刻画 舆情事件传播生命周期反映了舆情信息的不同发展阶段与舆情信息的生命力，舆情信息的生命力又反映了舆情信息所含内容的传播有效性。虽然在舆情研究领域中，由于分析案例的传播生命周期演变具体情况不同，会有不同的划分结果，但是根据生命周期理论，一般性事件的生命周期可被切分为萌芽期、成长期、衰退期、平稳期四个阶段，本研究将以这四个阶段为基础制定案例研究中传播周期阶段划分的规则。

3.1.3 基于 LDA 模型的不同传播者的主题挖掘与观点识别 本研究从两个层次对传播周期各个窗口期的舆情进行主题语义挖掘研究。第一层次：①按照生命周期划分的不同窗口期对语料库实现 LDA 主题抽取，根据舆情事件各阶段抽取的特征词进行归纳。②研究结合采集的语料语境对归纳的特征词进行主流词组的语义标注，并解读传播周期各阶段语义的意义。第二层次：根据标注和解读的语义词组，对传播周期各阶段的的主题进行观点识别。

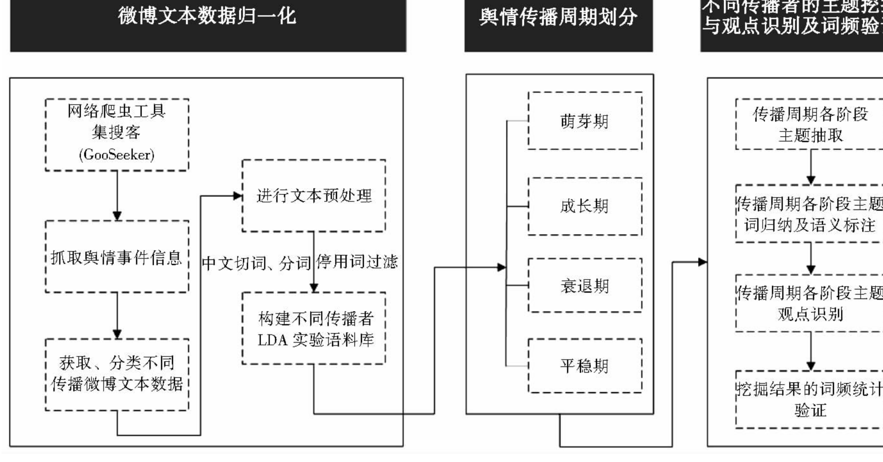
3.1.4 词频统计验证 应用相关工具或软件抽取分析语料的高频词，去除噪音、无意义的词汇，并进行词频高低的排序。

根据上述步骤描述的内容，本文提出基于不同传播者的微博舆情主题挖掘与观点识别的研究流程设计见图 1（图 1 中包含了本文具体预设的研究方法、工具以及各个环节的关键问题）。

3.2 基于生命周期理论的微博传播周期阶段划分的规则

在生命周期理论的划分中，通常学者们会根据具体的应用场景划分为 3 个或 4 个阶段，在此，本文对微博舆情传播周期进行一般意义上的生命周期阶段的划分，并提出划分规则：

（1）萌芽期：微博舆情发文、评论量较少，传播增量



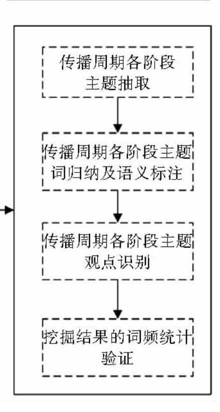


图 1 研究流程设计

几乎为零甚至为负增长,言语匮乏单一,话题种类较少,但不断有新的词语出现,表明此时传播情况处于萌芽期。

(2)成长期:新浪微博呈现爆发式增长,出现指数增长形式,发文、评论量呈现增长状态,新的留言、评论不断增加。与此同时,舆情话题数量激增,发文和评论量随时间呈现激增的曲线形式说明传播的迅速到达一个峰值,表明了舆情传播在此阶段进入了超速爆发的阶段。

(3)衰退期:舆情事件发文、评论量呈现迅速下降趋势,传播量增长率递减,词语增长率有可能为负,话题数基本不更新,表明该阶段舆情传播正在退出热门话题,传播量明显下降。

(4)平稳期:舆情事件的传播在经过衰退期以后,每日的传播量进入相对稳定的时期。该阶段传播量的增长率几乎为零,并且传播能量经过萌芽期、成长期、衰退期后,已经进入了传播群体情感宣泄、思想表达的消极期。然而,此时的传播情况还可能会出现两种趋势,一是该主题传播量递减,没有新的语言动向,传播量维持在一定的稳定水平。二是在原有主题的基础上,由于新闻消息的披露,一些词语量呈正增长趋势,舆情信息传播量有所递增。表明该事件衍生出了新的热点话题,新的舆情也即将爆发。

3.3 基于 LDA 模型的不同传播者的热点主题模型构建

根据划分的传播周期,本文结合 LDA 主题模型构建微博舆情事件环境下的热点主题研究模型。构建模型的主要维度有:时间周期、微博用户、微博内容、主题挖掘与观点识别,维度描述如下:

(1)时间周期维度:舆情事件的传播是由时间流和信息组成。根据舆情信息传播周期的发展趋势,本文以 1 天为时间粒度,将舆情进行生命周期阶段的划分。

(2)微博用户维度:用户是舆情发布和生产的主体。本研究基于用户发布、用户评论两个层面的信息源进行传播主题的挖掘。

(3)微博内容维度:微博用户的发布、评论信息中隐含着的用户对舆情事件的情感倾向、思想、观点、意见等主观态度的内容。

(4)主题挖掘与观点识别维度:基于信息采集和 LDA 主题抽取获取舆情主题,并将隐藏在微博话语中的主题进行语义标注。根据主题的语义分析结果,再将观点高度概括和总结出来。

然而在上述维度的描述中,微博数据仍然存在很多特殊性:①在时间周期维度层面中,舆情事件所持的主题往往在很短的时间内就发生变化,即同一阶段内,出现多个主题。②在用户维度层面中,同一用户可能在不同的时期发布不同的内容,或者发布相似的内容。③在内容维度层面中,一条微博内容可能表达了几种观点,也有可能是几条微博都属于同一类观点。所以,研究根据微博数据的特征和特殊性,构建由时间周期、微博用户、内容三个维度因素融合影响下的主题挖掘模型,主题挖掘与观点识别的维度视为这三个因素作用的结果,研究模型见图 2。本文力求构建不同传播者的热点主题研究模型能够反映出舆情传播的势态,同时能为舆情管理控制的实时监测和危机应对提供有参考价值的结论。

4 实证分析

4.1 研究数据采集与处理

4.1.1 数据采集及基本情况描述 本研究利用网络爬虫工具集搜客 GooSeeker 对“8.12 天津爆炸事件”进行数据采集。研究抓取数据包括:微博发布内容、微博评论内容、微博发布者 ID、微博评论者 ID、微博 ID 等。采集时间为 2015 年 8 月 12 日到 2015 年 9 月 13 日。图 3 是统计的采集微博发布和评论数据总的传播量趋势。

根据“8.12 天津爆炸事件”传播量趋势,本文对所研究的舆情案例进行舆情传播周期的切分, 周期分为

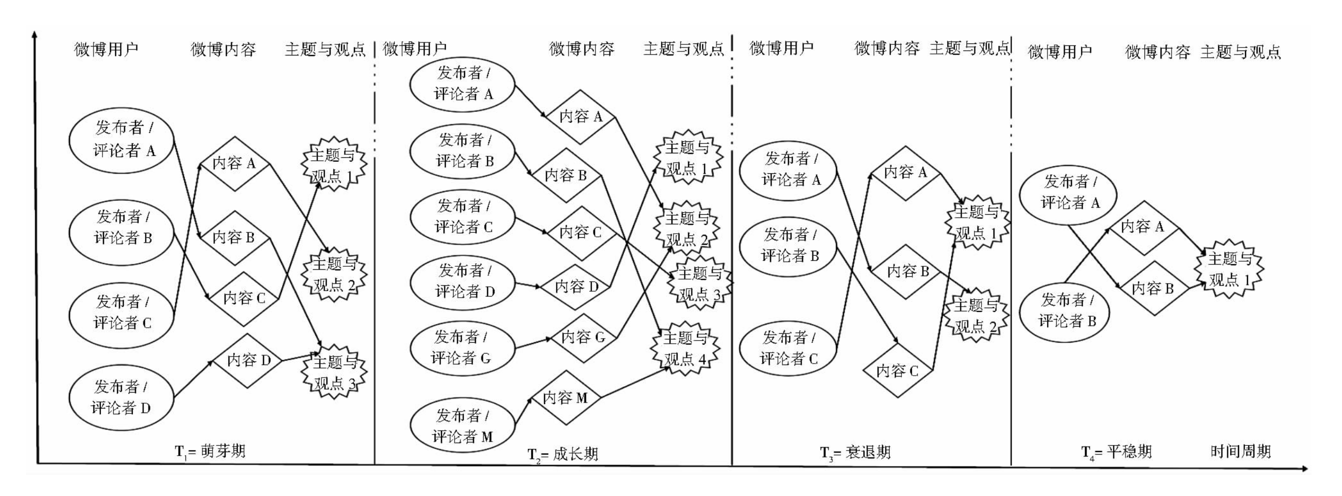


图2　不同传播者的热点主题研究模型

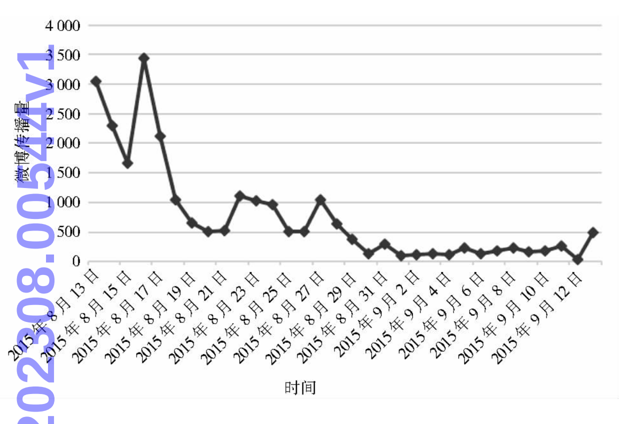


图3　“8.12 天津爆炸事件”每日传播量

3个阶段:萌芽期(8月12日到8月15日)、爆发期(8月16日到8月31日)、平稳期(9月1日到9月13日)。在划分阶段中,8月16日当日传播量出现激增,并且到达了传播的最高峰值,是成长期显著的传播特征;从8月17日到31日,传播总量呈现迅速下降的衰退趋势,是衰退期显著的传播特征;由于在舆情传播中,这两种传播态势都存在于事件的发酵高温期,如果分阶段计量分析,不利于主题的挖掘,所以将两个阶段合并定义为爆发阶段。其余阶段的传播趋势均符合舆情生命周期的划分规则。

4.1.2　文本处理实验步骤　根据微博信息发布性质的不同,研究将传播者定义为微博发布者和微博评论者。微博发布者包括官媒、大众媒体等公共媒体,微博评论者包括多数普通用户等信息接收者。针对不同类型的传播者,研究建立两类语料库,即微博发布者语料库和微博评论者语料库。根据研究的需求,本文选择了微博发布的长博文和用户的短评论作为初始的语料。由于采集的“8.12 天津爆炸”事件的微博文本中黑话、痞话、片段话语、谣言随处可见,这些词汇表达激烈、隐晦且书写形式相当不规范,使得初始语料的文本语言杂乱无章。本文考虑到文本分析的效率性,对微博所表达的观点、语义结合博文本内容进行总结、归类,在进行人工噪音处理后建立了规范的基础语料库。

研究再利用 jieba 分词工具包实现中文分词、去除停用词等自然语言处理,除去如“有点”“木子”“感觉”“无法”“呵呵”等无实际意义的词语,得到了整齐结构化的数据,以此作为实验所用的语料。然后,再基于开源 gensim 包实现 LDA 主题模型的参数训练。本研究的 LDA 参数设置参考了相关文献的方法[22],设置迭代次数为 2 000 次,超参数设置 $\alpha = 0.01, \beta = 0.05, k = 10$。参数值确定后,输入语料文件,运行 LDA 建模程序。LDA 主题抽取后获得两个重要结果文档。一个是主题分布文档,该文档用来计算主题强度;另一个是特征词分布文档,是每个主题下的特征分布的词项及概率。

4.1.3　主题抽取结果展示　研究分别对两个实验语料库进行萌芽期、爆发期、平稳期的 LDA 主题抽取,得出每个阶段的 10 个抽取主题及该主题下的相关词项。由于篇幅限制,本文选取了抽取的部分主题文档作为结果展示,见表1和表2。

4.2　微博舆情传播周期中不同传播者的主题语义挖掘

为了更准确地挖掘、解读主题语义,本研究选取传播周期各阶段强度前三的热点主题进行分析,且择优选取热点主题下概率较高的十项特征词进行主题解读。那么,热点主题特征词归纳结果见表3。

根据以上表3,本文结合语料库对特征词进行主流语义的标注,以得到各阶段热点主题的关键词组,结果见表4。

表 1　萌芽期微博发布者主题抽取结果

Topic1		Topic2		Topic3		Topic4	
0.009 885	公司	0.011 977	机构	0.008 522	居民	0.008 157	现场
0.008 627	现场	0.011 897	事故	0.008 478	距离	0.008 157	事故
0.007 285	企业	0.011 602	全面	0.008 395	悲剧	0.008 157	消息
0.006 804	人员	0.011 266	物品	0.008 300	发生爆炸	0.008 157	发生爆炸
0.006 727	事故	0.010 929	范围	0.008 258	规定	0.008 157	码头
0.006 326	地点	0.010 786	产品	0.008 195	信息	0.008 157	冲击波
0.006 313	发生爆炸	0.010 538	公司	0.008 169	规划	0.008 157	感觉
0.006 160	物流	0.007 659	标准	0.008 162	居民区	0.005 447	集装箱
0.006 050	仓库	0.007 659	居民	0.008 155	时间	0.005 447	医院
0.006 046	港口	0.007 659	平均价格	0.008 150	区域	0.005 447	仓库
……	…….	……	……	……	…….	…….	……

表 2　萌芽期微博评论者主题抽取结果

Topic1		Topic2		Topic3		Topic4	
0.001 367	朋友圈	0.001 965	逝者	0.001 130	少女	0.001 879	朋友
0.001 356	照片	0.001 821	感觉	0.001 124	谣言	0.001 857	消防员
0.001 352	回家	0.001 750	用户	0.001 116	部队	0.001 802	现场
0.001 326	少女	0.001 718	灾难	0.001 115	化学	0.001 781	学校
0.001 313	孩子	0.001 635	医院	0.001 108	人数	0.001 760	评论
0.001 302	国际	0.001 628	卫视	0.001 104	人祸	0.001 750	谣言
0.001 298	消防员	0.001 558	家人	0.001 104	小心	0.001 740	事故
0.001 284	老百姓	0.001 544	回家	0.001 095	仓库	0.001 715	事情
0.001 276	新闻	0.001 511	新闻	0.001 076	信息	0.001 690	名字
0.001 268	同学	0.001 461	信息	0.001 039	灾难	0.001 669	同胞
……	…….	……	……	……	…….	…….	……

表 3　微博舆情传播周期各阶段热点主题特征词

传播者类别	传播阶段	热点主题特征词
微博发布者	萌芽期	Topic(5.97):现场(0.008 627) 人员(0.006 805) 事故(0.006 727) 地点(0.006 326) 爆炸(0.006 312) 物流(0.006 16) 仓库(0.006 05) 港口(0.006 046) 负责人(0.005 776) 集装箱(0.005 764)
	爆发期	Topic(4.82):行业(0.009 15) 部门 (0.007 622)有限公司 (0.007 564)规定(0.007 553)企业(0.006 909) 危化品(0.006 88) 依法(0.006 514)专案组(0.006 031)立案侦查(0.006 031) 公安机关(0.006 031) 强制措施(0.006 031)
	平稳期	Topic(1.001): 事态(0.00623) 保险公司(0.006189) 隐性(0.006) 巨头(0.005 978) 根本性(0.005 953) 专业化(0.005 946) 爆炸(0.005 923) 航运(0.005 918) 进出港(0.005 888) 制造商(0.005 845)
微博评论者	萌芽期	Topic(7.63):事故(0.002 624) 评论(0.002 368) 医院(0.002 096) 视频(0.002 078) 消息(0.001 984) 新闻(0.001 945) 消防(0.001 842) 事情(0.001 788) 关心(0.001 769)能量(0.001 765)蘑菇云(0.001 758)
	爆发期	Topic(6.008):灾难(0.001 588) 责任(0.001 479)战士(0.001 473) 买单(0.001 466)事件(0.001 437) 心痛(0.001 433) 社会(0.001 404) 交代(0.001 4) 百姓(0.001 391) 消防员(0.001 372)
	平稳期	Topic(3.55): 群众(0.003 069)记者(0.003 042) 法律(0.002 99) 力度(0.002 978) 民众(0.002 969) 企业(0.002 958) 政府(0.002 953) 老百姓(0.002 945) 制裁(0.002 945) 时间(0.002 913)

表 4　微博舆情传播周期各阶段热点主题的关键词组

微博发布者	萌芽期	Topic:现场火光 受伤人员 天津爆炸事故 爆炸地点 瑞海物流 危险品仓库 港口管理 企业负责人 集装箱内易燃易爆物品
	爆发期	Topic:化工行业 安监部门 瑞海国际物流有限公司 安全距离规定 危化品经营企业 依法追究 检察调查专案组 立案侦查 天津公安机关 采取强制措施
	稳期	Topic:跟踪事态 保险公司赔偿 隐性影响 化工巨头 治理体系根本性 专业化治理 天津爆炸 航运公司 停止进出港 汽车制造商
微博评论者	萌芽期	Topic:爆炸事故 写评论 送往医院 看视频 官方消息 刷新闻 消防人员 这种事情 正能量 看见蘑菇云
	爆发期	Topic:灾难面前 承担责任 英雄战士 来买单 爆炸事件 太心痛 社会责任 给交代 百姓们 为消防员祈福
	平稳期	Topic:受灾群众 新闻记者 法律制裁 安全监管力度 天津民众 大企业 相信政府 保障老百姓 第一时间

（1）不同传播者的语言特征解读。观察表4整体内容，发现微博发布者语言较为正规化、专业化、多使用陈述性的名词，如："危险品仓库""安全距离规定""立案侦查""治理体系根本性"等；微博评论者语言较为日常化、口语化、多使用动词，如："写评论""刷消息""看见蘑菇云"等。

（2）不同传播者传播周期各阶段主题语义分析。在微博发布者主题挖掘方面：①萌芽期中，微博发布的热点主题是"事故报道"，该主题包含了与热议事故相关的主要信息，如集装箱、危险品仓库等；以及与事故相关的关键人物，如企业负责人、受伤人员。②爆发期中，微博发布的话题与前一阶段相比有了转换。由原先的事故相关信息的发布、报道，转变成了"事故追查"，新的热点主题体现了发布者对事故信息的深度挖掘。从该主题下的关键词组组成的逻辑可以看出，信息披露了企业爆炸出现的原因，如违背化工行业标准，并且传递了事故追责、调查、立案的动向消息。③平稳期中，微博发布的热点主题是"灾后盘点"，由关键词组可以看出，内容涉及经济赔偿、损失的盘点以及提出企业经营管理问题的治理要求。

在微博评论者主题挖掘方面：①萌芽期中，微博用户评论的热点主题是"事故讨论"，从该主题可以了解到萌芽阶段用户通过刷新闻，看视频等方式接收消息；用户还探讨了医疗情况、浓烟势态、并呼吁正能量言论。②爆发期中，微博用户评论的热点主题是"事故情感"，该主题下的来买单、太心痛、给交代、祈福等词都表征了用户在该时间段的悲痛、气愤、无奈等的情感和追求真相的心理状态。③平稳期中，微博用户评论的热点主题是"灾后安排"，该主题的关键词组中出现的相信政府、保障百姓、安全监管等词汇，说明了老百姓相信政府会出台有效措施、政策安排灾后民众的民生问题，以及关注对事故责任人的法律制裁的心理诉求。

由以上分析可知，微博发布者与评论者的萌芽期热点主题分别为"事故报道"与"事故讨论"，爆发期分别为"事故追查"和"事故情感"，平稳期分别为"灾后盘点"和"灾后安排"。各阶段主题对比发现，政府官方、新闻媒体发布的内容基本上属于事件消息的披露，用户发布的内容基本上是表达观点、情感和心理诉求等，且微博发布者主题和评论者主题具有较大的差异性。

4.3 基于主题挖掘结果的不同传播者的热点主题观点识别

本文根据主题挖掘结果，结合微博语料的具体语境选出代表性微博，并进一步进行人工识别和主题观点的总结。下图4是本文整理的每阶段传播的热点主题观点及传播演化情况：

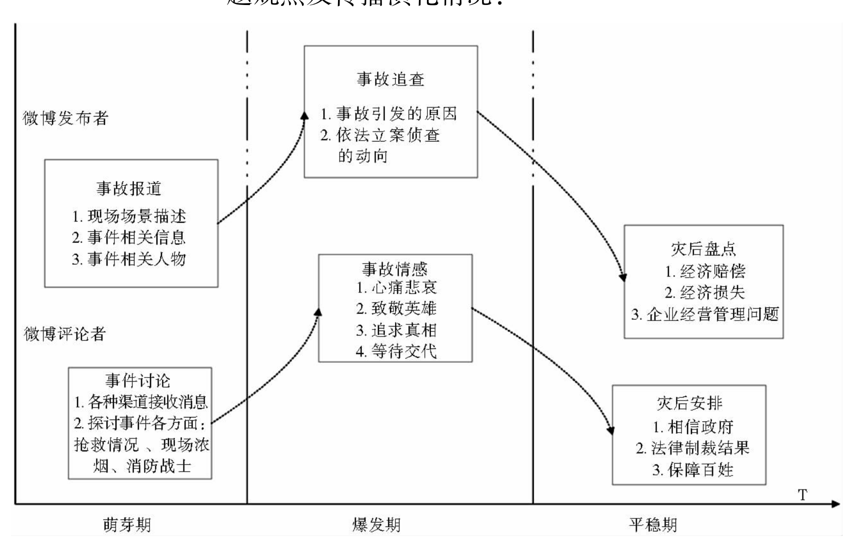


图4　不同传播者的热点主题观点及演化趋势

基于以上图4发现，在事件传播的过程中，微博发布者和微博评论者的传播话题的主题观点各有特征和偏向；同时，图4也展示了本研究舆情事件传播主题的结构和脉络。如微博发布方面主题谈论的是"事故报道""事故追查""灾后盘点"，而用户评论方面则谈论的是"事件讨论""情绪表达""灾后安排"。由传播周期主题演化脉络可以看出，研究基于案例挖掘出的热点主题符合事件发展的逻辑，和发布、评论对象的说话逻辑。由此，说明了本文提出的结合生命周期理论和LDA模型的舆情事件热点主题分析系统具有科学性和有效性。

4.4 研究结果的验证

为了对主题挖掘结论的有效性以及基于LDA和生命周期组合的主题分析体系的可靠性进行验证，本研究以词频秩序分析法统计了"8.12天津爆炸"事件传播周期中的舆情内容Top10的高频词和词频，统计结果见表5。

由表5可知，微博舆情传播周期Top10高频关键词代表了全局和每个阶段的10个研究主题，即研究主题从单个关键词来定义。如全局高频关键词中，"爆炸"这个关键词出现了417次，说明"爆炸"是最受关注的主题，但该主题下所包含的深层次的语义信息无法得知，而LDA方法所得到的结果是一个主题以及该

表 5 微博舆情传播周期 Top10 高频词

全局		萌芽期		爆发期		平稳期	
微博发布者	微博评论者	微博发布者	微博评论者	微博发布者	微博评论者	微博发布者	微博评论者
爆炸 417	新闻 64	爆炸 185	新闻 33	天津 160	社会 22	爆炸 101	老百姓 10
天津 367	逝者 61	天津 124	逝者 29	爆炸 131	灾难 22	天津 83	社会 10
事故 247	感觉 59	事故 84	视频 28	滨海 117	消防员 22	事故 48	灾难 10
滨海 199	事故 58	现场 82	消息 28	事故 115	责任 22	天津港 40	政府 10
安全 182	灾难 57	安全 78	朋友 27	安全 101	消防员 22	新区 39	爆炸事件 9
天津港 156	事情 56	消防 62	感觉 27	天津港 78	时间 22	滨海 37	新闻 9
新区 138	评论 55	人员 54	医院 27	企业 61	新闻 22	业主 33	企业 9
现场 135	消防员 54	距离 52	事故 27	新区 60	交代 22	居民 23	责任人 8
企业 114	地方 54	危险 47	仓库 26	天津市 55	事故 22	损失 23	牺牲者 8
消防 112	国家 54	滨海 45	消防 24	房屋 43	问责 21	天津市 22	制裁 8

主题下相关词项。将表 5 与抽取的主题关键词结果表进行对照分析，从表 5 中的全局高频词列可以看出，微博发布者和评论者抽取的主题词基本涵盖高频词，这表征基于 LDA 抽取的词项具有准确性。

在通过同样的方法观察对比各阶段高频关键词的结果，如见表 5 的第三列代表的是萌芽阶段微博发布者的高频词，以及表 1 萌芽期微博发布者主题抽取的结果。对比可知，LDA 的特征词也基本涵盖了高频词项。表 5 中的第五列关键词显示：天津、企业、新区、天津市、房屋等高频词，研究发现这些高频词之间的语义意境跨度较大，词项间的关系无从得知，按照频次高低排列在一起也很难识别出语义信息。而 LDA 得出的主题抽取结果，如表 3 中微博发布者萌芽阶段显示的现场、人员、事故、地点、爆炸、物流、仓库、港口等词是包含在“事故报道”这个主题下的特征词，而这些词项的集合也体现出了该主题的意义，并且由 LDA 的特征词项还可以看出，LDA 抽取的结果能够读取的信息不仅具体且更为丰富多元。

5 结论

本研究以微博“8.12 天津爆炸事件”为例，对不同传播者的发表内容进行 LDA 模型和生命周期理论结合的主题挖掘与观点识别分析。研究得出了一些有意义的结论，总结如下：

在理论意义方面，本文将 LDA 模型和生命周期理论结合，实现了方法论与理论的融合应用。将构建的 LDA 热点主题挖掘模型，应用到特定事件的案例中，通过挖掘模型的维度分析、层次分析、角色分析，研判出分布在传播当中有关联性的、代表性的、重要的词语，以及不同传播者热点主题的异同，大大提高语料信息的解释性。但由总体的主题分析而言，政府官媒、大众媒体和用户群体话题有着显著的差异性，显然微博平台中公众用户孕育出了自己的个性话题。而生命周期理论，是针对于传播主题的刻画，达到了展现宏观事态结构的功能；再从微观方面来讲，本文挖掘出了时间粒度下的主题内容及相关信息，并展示出每阶段的影响力话题，深化了社会舆情研究，为决策提供更多的信息。在现实意义方面，本文具有实时监测的意义。本文识别出的主题观点，能够了解舆情势态变迁、演化过程、公众的思想观念转变等情况，且本文借助生命周期的划分观察，起到了舆情监测的目的。

然而，本文的研究仍然是有一定的局限和困境的。本文建立的研究模型，能够有效的挖掘舆情传播周期的热点主题，但方法仍有待改进。LDA 主题模型抽取的特征词也不能像人工那样完整的解读一句话的意义，能够挖掘的结论仅限于在集中词项聚类下表达的主题的意义，这也是 LDA 方法解读文本的局限性。因此，如何挖掘出言论中更多的有效信息，仍是需要进一步解决的问题。在未来研究的发展方向上，本文考虑舆情传播主题与主题、主题与事件、主题与媒介、应用情境等元素之间的关联；在研究方法上，加强和深化定性分析，使得研究更加具有社会科学性；同时也力求未来的研究通过这些新增视角、元素、方法等能够更加深入地刻画舆情事件隐藏主题的遗传和变异，也更加清晰地展现事件发展的脉络和趋势。那么，如何从主题关联视角进行舆情事件的语义挖掘，将是今后研究需要进行的重点。

参考文献：

[1] 陈晓美，高铖，关心惠. 网络舆情观点提取的 LDA 主题模型方法[J]. 图书情报工作，2015，59(21)：21 – 26.
[2] 张寿华，刘振鹏. 网络舆情热点话题聚类方法研究[J]. 小型微型计算机系统，2013，34(3)：471 – 474.
[3] 李磊，刘继，张竑魁. 基于共现分析的网络舆情话题发现及态势演化研究[J]. 情报科学，2016，34(1)：44 – 47，57.
[4] 钱爱兵. 基于主题的网络舆情分析模型及其实现[J]. 现代图

书情报技术,2008(4):49 - 55.
[5] 梁晓贺,田儒雅,吴蕾,等.基于超网络的微博舆情主题挖掘方法[J].情报理论与实践,2017,40(10):100 - 105.
[6] LI N, WU D D. Using text mining and sentiment analysis for online forums hotspot detection and forecast[J]. Decision support systems, 2010, 48(2): 354 - 368.
[7] SU L Y F, CACCIATORE M A, LIANG X, et al. Analyzing public sentiments online: combining human-and computer-based content analysis[J]. Information, communication & society, 2017, 20(3): 406 - 427.
[8] 丁晟春,龚思兰,周文杰,等.基于知识库和主题爬虫的南海舆情实时监测研究[J].情报杂志,2016,35(5):32 - 37.
[9] 张瑜,李兵,刘晨玥.面向主题的微博热门话题舆情监测研究——以"北京单双号限行常态化"舆情分析为例[J].中文信息学报,2015,29(5):143 - 151,159.
[10] 安璐,吴林.融合主题与情感特征的突发事件微博舆情演化分析[J].图书情报工作,2017,61(15):120 - 129.
[11] ZHAO J, DONG L, WU J, et al. Moodlens: an emoticon-based sentiment analysis system for chinese tweets[C]//Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2012: 1528 - 1531.
[12] MEI Q, LING X, WONDRA M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs[C]//Proceedings of the 16th international conference on World Wide Web. New York: ACM,2007: 171 - 180.
[13] 关鹏,王曰芬.学科领域生命周期中作者研究兴趣演化分析[J].图书情报工作,2016,60(19):116 - 124.
[14] 唐晓波,向坤.基于LDA模型和微博热度的热点挖掘[J].图书情报工作, 2014, 58(5):58 - 63.
[15] 林萍,黄卫东.基于LDA模型的网络舆情事件话题演化分析[J].情报杂志,2013,32(12):26 - 30.
[16] ZHAO W X, JIANG J, WENG J, et al. Comparing twitter and traditional media using topic models[C]//European conference on information retrieval. Berlin,Heidelberg:Springer-Verlag,2011: 338 - 349.
[17] PENNACCHIOTTI M, GURUMURTHY S. Investigating topic models for social media user recommendation[C] //Proceedings of the 20th international conference companion on World wide web. New York: ACM, 2011: 101 - 102.
[18] 安璐,杜廷尧,余传明,等.突发公共卫生事件的微博主题演化模式和时序趋势——以Twitter和Weibo的埃博拉微博为例[J].情报资料工作, 2016(5):44 - 52.
[19] 陈福集,黄江玲.基于演化博弈的网络舆情热点话题传播模型研究[J].情报科学,2015,33(11):74 - 78.
[20] 张思龙.微博热点话题预判技术研究[D].郑州:解放军信息工程大学,2013.
[21] MEI Q, LIU C, SU H, et al. A probabilistic approach to spatiotemporal theme pattern mining on weblogs[C] //Proceedings of the 15th international conference on World Wide Web. New York: ACM, 2006:533 - 542.
[22] 关鹏,王曰芬.基于LDA主题模型和生命周期理论的科学文献主题挖掘[J].情报学报,2015,34(3):286 - 299.

作者贡献说明:

廖海涵:提出研究思路,整理、分析数据,撰写论文;
王曰芬:扩展研究思路,论文定稿;
关鹏:数据采集、处理。

Topic Mining and Viewpoint Recognition of Different Communicators in the Transmission Cycle of Micro-blog Public Opinion

Liao Haihan[1] Wang Yuefen[1,2] Guan Peng[1]
[1] School of economics and management, Nanjing University of Science and Technology, Nanjing 210094
[2] Social Public Safety Science and Technology Co-Innovation Center, Jiang Su Province, Nanjing 210094

Abstract: [**Purpose/significance**] This paper aims to explore the hot spot of public opinion and the main point view of the communication of different communicators in the transmission cycle of micro-blog public opinion and to discover the characteristics and laws of public opinion transmission, which can provide the basis for public opinion analysis and decision making. [**Method/process**] This study is based on the text data of a true public opinion event. It adopted life cycle theory and LDA method to design research process and construct research model, and researched topics of different communicators in micro-blog public opinion events, including topic extraction and semantic annotation, semantic analysis of different communicators at various stages, recognition and characterization of theme views of public opinion based on time dimension. [**Result/conclusion**] It is found that the research model proposed in this paper can excavate topic theme structure, view and characteristics of different communicators in the communication cycle of public opinion. And the words with actual meaning and irritating function are related, representative and important. At the same time, the conclusion also found a hot topic in the mass media or the official micro-blog is totally different from micro-blog users.

Keywords: micro-blog public opinion　different communicators　topic mining　view identification　life cycle theory　LDA theme model